Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

Course No: EE 3122
Sessional on Numerical Methods & Statistics

Experiment No. 5

Name of the Experiment: Curve Fitting and Correlation

Objectives:

[1] To implement linear correlation and Rank correlation
[2] To implement least square regression (line fitting) and least square parabola or polynomial fitting

Theory/Introduction:

- The direction and strength of the relationship between the paired *x* and *y* values in a sample can be expressed by means of a correlation coefficient "r", which is mathematically defined as:

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{SCP}{\sqrt{(SSX)(SSY)}}$$

- The sum of cross products of deviations

$$SCP = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

- The sum of squared deviations for X

$$SSX = \sum (y_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

- The sum of squared deviations for Y

$$SSY = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

- It is also called **Pearson's correlation** or **product moment correlation coefficient**. It measures the nature and strength between two variables of the quantitative type. The sign of r denotes the nature of association while the value of r denotes the strength of association.

- Pearson's correlation of coefficient formula $r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}$, where the covariance is given by $\text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ and $S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$, and $S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$

Rank Correlation

Spearman rank correlation: Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.
The following formula is used to calculate the Spearman rank correlation:

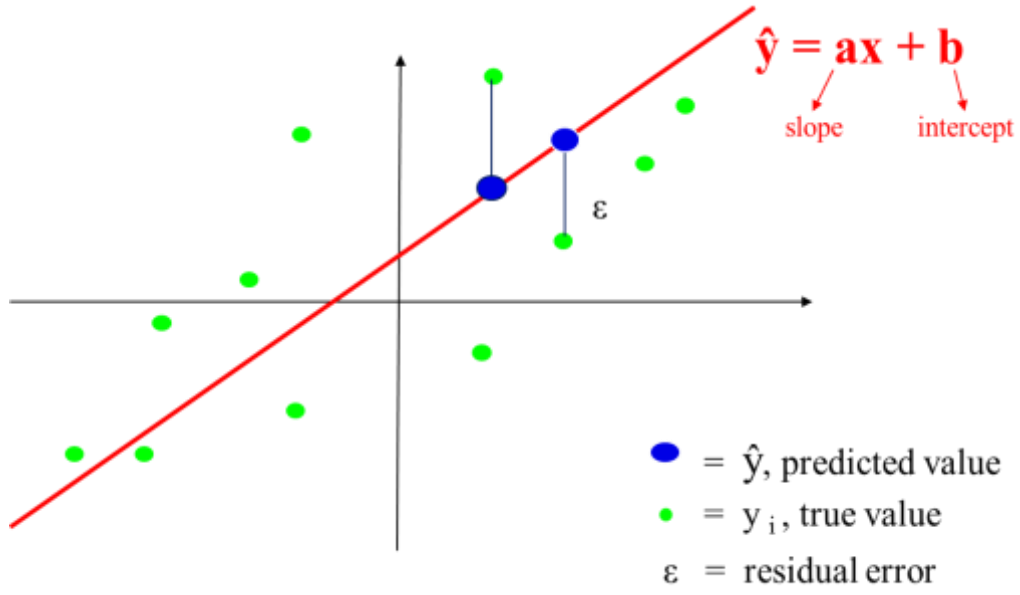
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ= Spearman rank correlation
d_i= the difference between the ranks of corresponding variables
n= number of observations

Fitting Data to a Straight Line:

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x
- This will be the line that minimises distance between data and fitted line, i.e. the residuals



- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)
- Model line: $\hat{y} = ax + b$, a = slope, b = intercept
- Actual line: $y = ax + b + \varepsilon$
- Residual (ε) $= y - \hat{y}$
- Sum of squares of residuals = $\Sigma (y - \hat{y})^2$
- we must find values of a and b that minimize $\Sigma (y - \hat{y})^2$
- The value of a and b is given by $a = \frac{rs_y}{s_x} = \frac{\text{cov}(x,y)}{s_x^2}$,
- Now, $\bar{y} = a\bar{x} + b$ and then $b = \bar{y} - a\bar{x}$

Algorithm:

[r, rho] = corr (xdata, ydata, Mode)

Develop the C++/Matlab code for fining the correlation coefficient, r and rank correlation coefficient, ρ (rho). Here, mode means either r or rho. Xdata and ydata represents the x-axis data and y-axis data respectively.
Steps: Read x-data from file
Read y-data from file
Find the standard deviation of x
Find the standard deviation of y
Find the covariance, cov(x, y) between x and y
From above value you can find the value of r
Using r, s_x, and s_y, you can calculate a
Using above values, you can estimate b
Now you can get the actual equation of line.
Plot the actual line and the scatter plot using x- and y-values
Sort x-data and convert into rank
Sort y-data and convert into rank
Find $\Sigma\ (rankx - ranky)^2$, then calculate the rho
end

Performance:

(i) Find the correlation, r between two sets of data. You need to use a single file or two files where the data are located.

```
xData = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10];
yData = [1.5, 2.6, 3.7, 4.8, 5.9, 7.0, 8.1, 9.2, 10.3, 11.4];
```

Use this data to fit a line y and find the error.

(ii) Find rank correlation for the above data.
(iii) Find the value of slope and intercept using the line y=ax + b *or* parabola
(iv) Plot the scatter plot using above data and line plot in a single graph.